An Overview of Corpus-Based Statistics-Oriented (CBSO) Techniques for Natural Language Processing (1996) [3 citations — 2 self]

by Keh-Yih Su , Tung-Hui Chiang , Jing-Shin Chang

Abstract:

A Corpus-Based Statistics-Oriented (CBSO) methodology, which is an attempt to avoid the drawbacks of traditional rule-based approaches and purely statistical approaches, is introduced in this paper. Rule-based approaches, with rules induced by human experts, had been the dominant paradigm in the natural language processing community. Such approaches, however, suffer from serious difficulties in knowledge acquisition in terms of cost and consistency. Therefore, it is very difficult for such systems to be scaled-up. Statistical methods, with the capability of automatically acquiring knowledge from corpora, are becoming more and more popular, in part, to amend the shortcomings of rule-based approaches. However, most simple statistical models, which adopt almost nothing from existing linguistic knowledge, often result in a large parameter space and, thus, require an unaffordably large training corpus for even well-justified linguistic phenomena. The corpus-based statistics-oriented (CBSO) ...

计算机科学1995Vol.22№.4

基于语料库和面向统计学的自然语言处理技术*)

周 强
(北京大学计算语言学研究所 北京100871)

摘 要 In this paper, some corpus-based, statistics-oriented natural language processing techniques, include: Shannon's noisy channel model and its applications, n-gram model, the methods to estimate and smooth arguments, preference-based parser and so on, were introduced. It was also discussed that how to use these techniques in the Chinese language processing.

关键词 Statistics-based processing, Corpus Linguistics, Natural language processing.

1 引 言

语料库语言学(Corpus Linguistics)是八十年代才崭露头角的一门新的计算语言学的分支学科。它研究机器可读的自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析,以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。语料库语言学研究的基础是机器可读的大容量语料库和一种易于实现的统计处理模型,两者相辅相成,缺一不可。从本质上讲,语料库语言学的研究采用的是一种基于统计的经验主义处理方法,与传统的基于规则的理性主义处理方法大相径庭。

早在1949年,Warren Weaver 就设想,可以利用信息论的编码思想,使用一种统计的方法来解决机器翻译问题。五十年代,经验主义处于鼎盛时期,统治了从心理学(行为主义)到电子工程(信息论)的广泛领域。那时候,不仅依据词的意义而且依据它们与其它词的共现情况对词进行分类,是语言学上的常规操作。但是,随着五十年代末到六十年代初一系列重大事件的发生,包括 Chomsky 在"句法结构"中对 n 元语法(n-gram)的批评和 Minsky 和 Papert 在"视觉感控器(Perceptrons)"中对神经网络的批评,经验主义便逐渐减退了。

近年来,计算机技术得到了飞速的发展,机器的存储量越来越大,运算速度越来越快,而价格却越来越便宜,使大容量的机器可读语料库的建设成为可能。仅仅在十几年以前,一百万词的 Brown 语料库还被认为是巨大的,但从此以后,出现了二千万词的 Birmingham 语料库。今天,许多地方都有了达到几亿甚至数十亿词的文本样例,一些新的、更好的统计语言模型也开始出现。而且,随着自然语言理解系统的不断实用化,知识获取问题已成为一个瓶颈,基于规则的 NLP 系统在处理大规模的非受限真实文本中遇到的种种困难,促使广大研究人员去探索和采用一种新的研究思想。所有这些因素,推动了基于语料库的经验主义研究方法成为目前 NLP 研究中的一个热点。本文主要根据笔者目前所掌握的一些资料,对基于语料库和面向统计学的经验主义处理技术作一个简要的介绍。

2 基于语料库和面向统计学的处理技术

在语料库语言学中,基于统计的处理技术是从语料库中获取各种所需要的知识的主要手段,它的基本思想是:i)使用语料库作为唯一的信息源,所有的知识(除了统计模型的构造方法)都是从语料库中获得的.ii)使用统计方法获取知识,知识在统计意义上被解释,所有参量都是通过统计处理从语料库中自动获得的。这种处理技术需要一定的概率论、信息论和数理统计的知识。下面简单地介绍一下其中的一些基本概念和术语:

1)概率 P(A) 表示在一个样本空间中,事件 A 发生的可能性。

2)条件概率 P(A|C) 表示在事件 C 发生的条件下,事件 A 发生的可能性。例如:给定一个特定的词 w,它在语料库中作名词 n 的概率为 P(n|w)。

3)联合概率 P(A,B) 表示事件 A 和 B 同时发生的可能性。例如,在语料库中,词 x 和词 y 同时出

*)本文得到自然科学基金资助。周 强 博士生,主攻方向为语料库处理、机器翻译、计算语言学。

现的概率为 P(x,y)。

4)贝叶斯计算模型　在概率论中，贝叶斯公式描述了通过一系列先验概率计算后验概率的一种方法，其具体定义为：

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n}P(B|A_j)P(A_j)},(i=1,2,\cdots,n)且$$

$$\sum_{i=1}^{n}P(A_i)=1$$

考虑其最简单的形式，则有：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}=\frac{P(B|A)P(A)}{P(B)}$$

此公式为解决语料库研究中大量的限制性对应问题提供了有力的支持。

5)平均值：$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i$　表示数列 $x_1,x_2,\cdots,x_N$ 的算术平均值。

6)方差：$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$　表示数列 $x_1,x_2,\cdots,x_N$ 相对平均值的离散程度。

7)熵：$H=-\sum_{i}P(x_i)\log P(x_i)$　是信息论中的一个重要概念，表示信源所具有的平均信息量的大小。

8)相关信息计算模型　在统计学中，相关信息(又称互信息)I(x;y)定义为：

$$I(x;y)=\log_2\left[\frac{P(x,y)}{P(x)\cdot P(y)}\right]$$

设 x,y 分别表示两个不同的单词，则 I(x;y)体现了词 x 和 y 信息的相关程度，即：若 $I(x,y)\gg 0$，则表明 x 与 y 是高度相关的；若 $I(x;y)=0$，则表明 x 与 y 是独立的；若 $I(x;y)\ll 0$，则表明 x 与 y 是互补分布的。

相关信息的计算对词相关(word association)和词共现(word co-occurrence)等信息的统计起着重要的作用。

3　噪声信道模型及其应用

3.1　噪声信道模型

香农的通信理论即是众所周知的信息论，最初是在 AT&T 贝尔实验室中为模型化沿着一条噪声信道(如：一条电话线)的通信问题而提出的。但作为一种抽象的理论模型，它在许多识别应用领域也得到了广泛的应用。

想象有这样一个噪声信道，它使一系列好的文本(I)进入信道后，以一系列讹误的文本(O)从另一端输出，即：I→噪声信道→O。一个自动过程怎样才能从一个讹误的输出 O 中恢复好的输入 I 呢？原则上，人们可以通过假设所有可能的输入 I，并且从中选取具有最高评分 P(I|O)的输入文本作为最有可能的输入 $\hat{I}$，符号化为：

$$\hat{I}=\arg\max_{I}P(I|O)=\arg\max_{I}P(I)P(O|I)$$

其中 argmax 表示寻找具有最大评分的参量。

先验概率 P(I)是 I 在信道的输入端出现的概率。例如：在语音识别中，它是说话人发出 I 的概率。但事实上，先验概率是得不到的，因此，我们需要构造一个先验概率的模型，如三元语法(3-gram)模型来模拟它。语言模型的参数可以通过计算大量文本样例上的不同统计数据而进行估计。

信道概率 P(O|I)是当 I 出现在输入端时 O 将在信道的输出端出现的概率。如果在某些合适的含义下，I 类似于 O，则此概率较大；反之，则较小。信道概率依赖于应用问题。例如：在语音识别中，单词"writer"的输出看起来可能类似于单词"rider"；而在字符识别中，"farm"则极有可能是"form"的输出。

3.2　在语言信息处理中的应用

(1)识别问题。在语音识别、光学字符识别(OCR)和自动拼写校对等大量的识别应用领域，噪声信道模型正越来越得到广泛的运用。这些识别问题都可以抽象为下面的模型：

W→噪声信道→Y

其中，W 是一串单词或字符。对于语音识别问题，Y 为一组声音信号；在 OCR 中，Y 为扫描得到的位图信息；而在拼写校对问题中，Y 则为一串可能有错的录入字符串。这样，问题的目标就归结于寻找这样的一个单词或字符串 $\hat{W}$，使：$\hat{W}=\arg\max_{W}P(W)P(Y|W)$。

(2)词类标注。目前的许多词类自动标注算法都是以香农的噪声信道模型为基础的。设有一串词类标记 C 出现在信道的输入端，并且由于某些奇怪的原因，它以一串单词的形式出现在信道的输出端：C→噪声信道→W。我们的工作就是要在给出 W 的情况下确定 C。最为可能的词类序列 $\hat{C}$ 可由下式给出：

$$\hat{C}=\arg\max_{C}P(C)P(W|C)$$

其中：P(C)和 P(W|C)可以利用从大规模标注文本中进行参数估计得到的一组语境概率 $P(c_i|c_{i-2}c_{i-1})$和一组词汇概率 $P(w_i|c_i)$进行简化计算而得到。在某种意义上，可以把这组语境概率看成一部语法，而把那组词汇概率看成一部词典。

(3)机器翻译。机器翻译(MT)研究究竟适合于

采用基于规则的理性主义方法还是基于统计的经验主义方法，是目前国际上争论的一个热点问题。对这两种方法都进行了一些研究和探索。Weaver(1949)第一次提出了一种对 MT 的信息论处理方法。五、六十年代，在乔治敦，这种经验主义方法也在一个系统中进行了实践，最终发展成人所共知的 SYSTRAN 系统。最近，MT 的大部分研究工作倾向于采用理性主义方法，但也有一些例外，如：基于实例的机器翻译(EBMT)研究。

IBM 的 P. F. Brown 等人的研究工作进一步发展了 Weaver 对 MT 信息论的处理方法，他们对法语翻译到英语的基本处理思路可以归结到香农的噪声信道模型中：E→噪声信道→F，这里的噪声信道可以想象为一种翻译机制。同以前一样，依据下列公式选择$\hat{E}$，可使错误几率达到最小：

$$\hat{E}=\arg\max_{E} P(E)P(F|E)$$

同样的，模型的参数估计可以利用大规模文本样例中得到的大量统计数据。其中先验概率 P(E)可以通过构造合适的英语语言模型加以估计，而信道概率 P(F|E)，则可以从由一个计算哪部分源文本对应哪部分目标文本的自动过程而建立了联结(alignment)的并行文本中进行估计。

(4)拼音汉字转换。自动转换问题是中文人机通讯中很关键的问题，它的解决对于人机自然语言交互通讯、汉字的键盘输入和汉语语音识别及合成都有重要意义。然而汉字的音字不一一对应，即一音多字、一字多音的现象，却给这个问题的解决带来了极大的困难，语料库语言学的发展，为研究者提供了一种新思路。

实际上，音字转换问题从抽象意义上看是一种对应问题，它非常类似于上面提到的识别问题，可以用噪声信道模型加以处理：W→噪声信道→E。一串汉字 W 经过信道后，以一串拼音 E 的形式输出，这样，问题的焦点就转化为寻找一个汉字串$\hat{W}$，使：

$$\hat{W}=\arg\max_{W} P(W)P(E|W)$$

利用这种方法的一些系统都取得了较好的转换效果。

4 统计模型构造和参量估计

在上面所提到的众多噪声信道应用问题中，如何计算先验概率 P(I)和信道概率 P(O|I)是研究的重点和难点所在，这需要根据不同的应用问题，选择并构造合适的统计语言模型，并利用从大规模文本样例中统计得到的大量数据来估计模型的参数。下面将简要地介绍模型构造和参量估计的常用方法。

4.1 统计模型的构造

对于先验概率，比较简单和常用的统计语言模型为 N 元语法(N-gram)模型。考虑单词串 $W=w_1,w_2,\cdots,w_n$，根据条件概率的定义，有：

$$P(W)=P(w_1w_2\cdots w_n)=\prod_{i=1}^{n}P(w_i|w_1\cdots w_{n-1})$$

其中 $P(w_n|w_1\cdots w_{n-1})$表示在给定历史信息 $w_1,w_2,\cdots,w_{n-1}$的条件下，选取词 w_n 的概率。这就是 N-gram 模型，并且所有信息组成了一条马尔可夫链。在实际应用中，为简化计算，往往只考虑一个或两个历史信息，形成了 bigram 模型($P(w_i|w_{i-1})$)和 trigram 模型($P(w_i|w_{i-2}w_{i-1})$)。

由于信道概率依赖于应用，因此需要根据不同的应用问题，选择合适的统计计算模型。下面通过两个具体的实例说明一下模型的构造方法。

(1)词性标注。对于单词串 $W=w_1,w_2,\cdots,w_n$ 和词类标记串 $C=c_1,c_2,\cdots,c_n$，假设每个词与词类标记的对应情况都是独立的，并且每个单词仅仅依赖于它自己的词类信息，于是可以得到如下的简化计算模型：

$$P(W|C)=P(w_1w_2\cdots w_n|c_1c_2\cdots c_n)=\prod_{i=1}^{n}P(w_i|c_i)$$

(2)机器翻译。考虑从英语到法语的单句翻译情况，可以发现，为把一句英语句子 $SE=we_1,we_2,\cdots,we_m$ 中的词 we_i 翻译为法语句子 $SF=wf_1,wf_2,\cdots wf_n$ 中的词 wf_j，一般可以采用下面三种方式：

a)直译(translation)。如在句子对(Jean aime Marie|John loves Mary)中，John 直译为 Jean，loves 直译为 aime，而 Mary 直译为 Marie。

b)繁殖(fertility)。有时英语单词可能翻译为多个法语词，如英语中的 not 在法语中常用 ne…pas 表示，则此词的繁殖率 f=2。但有时对英语句子中的某些词，在法语译句中可能没有任何词与之对应，这时可以认为，此英语单词的繁殖率 f=0。

c)变形(distortion)。由于语言使用习惯的不同，造成某些词群在位置关系上的变形。如：在英语中，修饰名词的形容词一般放在名词前，而在法语中，形容词常放在名词后。

对这三种方式进行抽象，就形成了如下的翻译模型：

$$P(SF|SE)=\prod_{i=1}^{m}\left[P(f_i|we_i)\cdot\prod_{j=1}^{f_i}P(wf_j|we_i)\right]\cdot\prod_{i,j}P(i|j)$$

其中 $P(f_i|we_i)$，$P(wf_j|we_i)$和 $P(i|j)$分别为繁殖概

率、直译概率和变形概率。它们都可以从建立了联结的英法双语语料库中进行参数训练而得到。

4.2 参量估计方法

(1)最大似然估计(MLE)。假设一个单词 w 在语料库中出现的概率 P(w)符合二项分布规律，则当语料库容量 N 足够大时，我们可以期望单词 w 将出现 N * P(w)次，从而得到 P(w)的估计值为：

$$P(w)=f(w)/N$$

其中 f(w)为单词 w 在语料库中出现的频度。这就是 MLE 估计方法。

这种方法简单而实用，在许多情况下都能得到比较合理的估计。但是，当数据不能很好地适应模型时，这种估计方法也可能出问题。研究表明，实词在语料库中的分布不能很好地符合二项分布规律，因为实词倾向于"突发性"地出现；由于某些文章风格因素的作用，虚词可能也会偏离二项分布。另外，由于统计数据的稀疏性，必然会出现一些语料库中不出现的情况，对此，MLE 方法将给出零概率的估计值，这给后续的计算处理带来了许多问题。所有这些不足，都需要寻找更精细的参数估计方法加以解决。

(2)数据稀疏性问题。我们可以通过表1中的数据来说明三元语法模型中统计数据的稀疏分布问题：

表1 n-gram 频度分布

(在 N=356,893,263个词的文本样例中)

频度	1-gram	2-gram	3-gram
1	36,789	8,045,024	53,737,350
2	20,269	2,065,469	9,228,958
3	13,123	970,434	3,653,791
>3	135,335	3,413,290	8,728,789
>0	205,516	14,494,217	75,349,888
≥0	260,741	6.799×10^{10}	1.773×10^{16}

表1中列出了在356,893,263个词的英语文本样例中出现的具有不同频度值的1-gram，2-gram，3-gram 参量的数目。所用的单词表包含了260,740个不同的单词，再加上一个未定义词词项，可将所有不在单词表中的词都映射到它上面。在所有可能的 $6,799\times10^{10}$个2-gram 参量中，只有14,494,217个参量真正在语料文本中出现，并且其中有8,045,024个只出现了一次。类似地，在所有可能的 $1,773\times10^{16}$个3-gram 参量中，只有75,349,888个参量真正出现在语料中，并且其中有53,737,350个仅仅出现了一次。从中，我们可以看到统计数据的稀疏性问题是很严重的。

显然，随着 n 的增大，n-gram 模型计算的精确度将不断增大，但由于训练文本数量的限制，参量估计的可靠性却在不断减低。为解决这个矛盾，就需要寻找新的技术以平滑统计数据。

(3)参数平滑方法

①插值估计。其基本处理思想为：将不同语言模型的参数估计通过插值公式组合起来，这样，当高级模型的参数估计比较可靠时，就利用这些更为精确的参数；反之，则退回到较低级的模型，使用那些不太精确但较为可靠的参数。令 $P^{(j)}(w_i|w_1^{i-1})$为第 j 个语言模型所决定的条件概率，则插值估计 $\hat{P}(w_i|w_1^{i-1})$可由下式给出：

$$\hat{P}(w_i|w_1^{i-1})=\sum_j\lambda_j(w_1^{i-1})P^{(j)}(w_i|w_1^{i-1})$$

给定 $P^{(j)}(w_i|w_1^{i-1})$的值，则 $\lambda_j(w_1^{i-1})$可用 EM 算法进行计算，且 $\sum_j\lambda_j(w_1^{i-1})=1$。考虑1-，2-，和3-gram 模型的插值估计，则有：

$$\hat{P}(w_i|w_{i-2}^{i-1})=\lambda_1P(w_i)+\lambda_2P(w_i|w_{i-1})+\lambda_3P(w_i|w_{i-2}w_{i-1})$$

Ney H.[1994]等介绍了一些更为复杂的非线形插值方法，这里就不再详述了。

②频度调整。其基本思路为，调整统计参量在语料库中出现的频度，以克服零概率问题。设某参量在语料库中出现 r 次，则据 MLE 方法，有 p=r/N。现令 r^*为 r 的调整频度，则此参量的概率就可估计为：

$$\hat{p}=r^*/N$$

为保证限制条件 $\sum p=1$，这个调整频度需满足：

$$\frac{\sum N_r\times r^*}{N}=1$$

其中 N_r为那些在语料库中频度出现 r 次的参量，即为频度 r 的频度，N 为语料库中总容量(总词数)。

最常见的频度调整方法为 Good-Turing 方法，它取 $r^*=(r+1)N_{r+1}/N_r$，类似的方法还有 held out 估计和 deleted 估计。Church K. 和 Gale, w. (1991)对这几种方法的处理性能进行了详细的分析和比较。

③其它常用方法。a)设置平伏常数：为所有零概率参量赋一个较小的数值 β(β≪1/N)。b)假定那些在语料库中没有出现的参量都出现一次，从而它们的概率值 p 就从0变为1/N。

5 基于优先的分析技术

自然语言中，词与词之间存在着许多优先组合

关系。词典编纂者使用术语：搭配(collocation)、共现(co-occurrence)和词关(lexis)来描述词对上的不同限制，一个典型的例子是 strong 和 powerful。Halliday([1966])注意到尽管 strong 和 powerful 具有类似的句法和语义，还是存在着各自更为适宜的不同语境(如：strong tea 和 powerfur computer)。心理语言学家也有一个类似的概念：词关联(word association)。两个经常引用的高度相关的例子是：bread/butter 和 doctor/nurse。心理学实验表明，对两个高度相关词的主题的反应比不相关词更为迅速。

这些限制或优先关系在计算语言学中很少讨论，因为它们通过传统的 NLP 技术，特别是基于规则的理性主义处理技术不能很好地获取。但是，建立一个能获取这些优先关系中的一部分的统计计算模型却不太困难，一个较常用的模型是信息论中的相关信息(mutual information)计算模型。

考虑词 x 和 y，相关信息 I(x,y)就反映了两个词之间的相关程度，其计算公式为：

$$I(x,y)=\log_2\left[\frac{P(x,y)}{P(x)\cdot P(y)}\right]$$

利用 MLE 方法估计 P(x)、P(y)，可以得到：

$$P(x)=\frac{f(x)}{N},P(y)=\frac{f(y)}{N}$$

而对联合概率 P(x,y)，则可以通过设置一个长度为 W 个词的观察窗口，移动这个窗口检索语料库的所有信息，统计词 x,y 在窗口中同时出现的次数 f(x,y)来加以估计，即：

$$P(x,y)=\frac{f(x,y)}{N}$$

显然，所选择的观察窗口的大小对统计结果的准确度有很大的影响。一般情况下，取 W=5，所提取的信息基本上可以满足要求了。

通过对词与词之间相关信息的计算，我们可以从语料库中提取许多有用的优先信息，如：名词和名词间紧密的语义联系(doctor/nurse)，形容词和名词组成的特定修饰关系(potent medicine 与 strong currency)，动词和名词的固定搭配(take a decision 而不用 make，pay attention 而不用 give)等。这些信息对于进行句法语义分析和自动排歧都很有用。

Hindle 和 Rooth 的研究就显示了共现统计数据在提高分析器的排歧能力上的作用。考虑这样一句英语句子：She(wanted|placed|put)the dress on the rack. 对于不同的动词，介词短语(on the rack)的连接方向是不一样。它可以修饰名词(对 wanted)，也可以作动词的宾语补足语(对 placed,put)，这就是英语句子分析中非常困难的介词短语连接(PP attachment)问题。Hindle 和 Rooth 的研究表明，一个分析器可以通过将动词—介词(want…on)间的相关信息值和宾语—介词(dress…on)间的相关信息值进行比较而选择合适的分析结果。另外，D. Magerman 利用相关信息计算模型来进行短语的自动划分，也取得了较好的效果。一些类似的研究还包括 Brent M.(1993)、Kobayashi Y. 等(1994)的工作。

实际上，基于统计优先的分析和基于规则的分析技术各有优势，因此理想的 NLP 模型应考虑把两者的能力结合起来。一个可能的方法是利用随机上下文无关语法(Stochastic Context Free Grammar，SCFG)。它为每个 CFG 规则 A→BC 赋一个概率值 P(A→BC)，有关的参数值可以利用 Inside-Outside 算法从语料库中训练得到。使用这种 SCFG 模型，一方面可以充分利用现有的 CFG 的成熟的分析技术；另一方面，通过引入统计概率，可以把大量的优先信息结合入分析器中，从而大大提高分析器的自动排歧能力。Bricoe T. 等(1993)、Tapanainen P. 等(1994)、Magerman D. 等(1990)在规则和统计相结合的分析技术研究方面进行了许多有益的探索。

6 结束语

本文简要地介绍了基于语料库和面向统计学的自然语言处理技术的基本内容，这些只是目前语料库语言学研究中的一小部分，其它许多有意思的研究课题，如：语料的平衡性问题、熵与语言模型的评估、对语言假设的解释数据分析(Explanation Data Analysis，EDA)、统计技术在词典编纂中的应用等，以后在条件成熟时，将另行撰文介绍。

从90年以来的数届重要的国际会议，包括 COLING，ACL，TMI 等，每届都有许多新的研究成果出现。面对汉语语料库语言学的研究，近几年来也出了许多研究成果，如：自动词性标注、自动分词研究、句法功能标注、语义信息标注、汉语音字转换、汉语语音识别等，但总的说来，发展速度并不是很快，规模也不太大。

笔者认为，目前汉语语料库研究的当务之急，是建立一个大规模的、经过多级加工处理的汉语语料库。这样的语料库至少应包含数百万、直至上千万词的覆盖各种题材的原始文本语料，然后经过自动切词，词性标注，句法结构分析和标注，语义标注等阶段的处理，形成一个具有不同处理层次、包含各种标注信息的语言知识库，从中可以提取大量有用的统计信息。当然，这是一项耗资巨大的工程项目，但它的建成，对于各种基于统计的汉语处理技术的发展，无疑会起巨大的推动作用。(参考文献共47篇略)